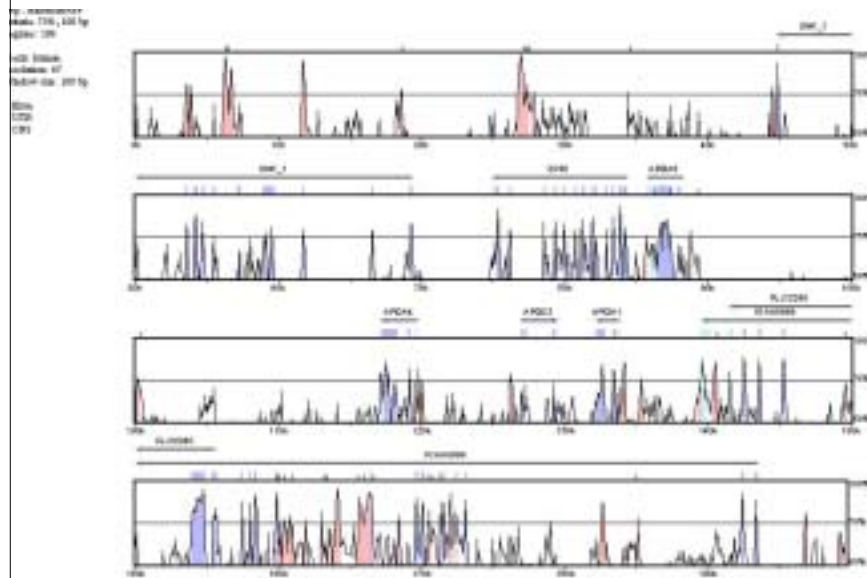


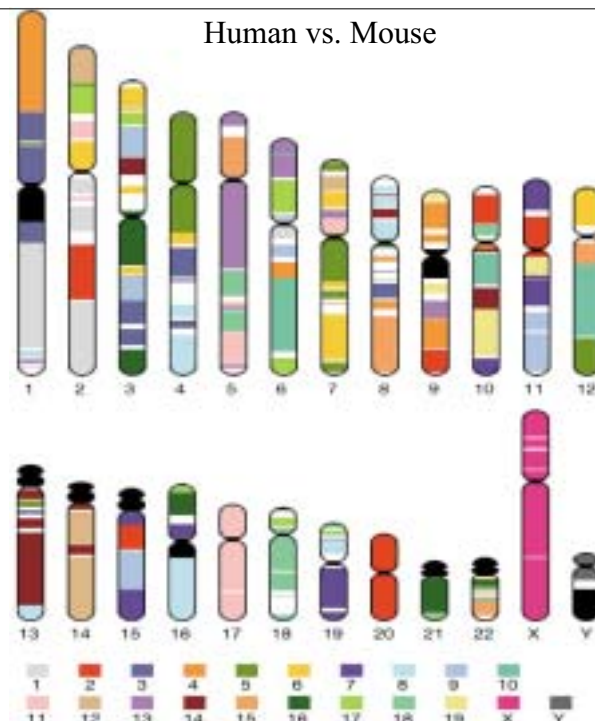
Comparison of Human and Mouse APOA1 Region



First thing you need:
Synteny:

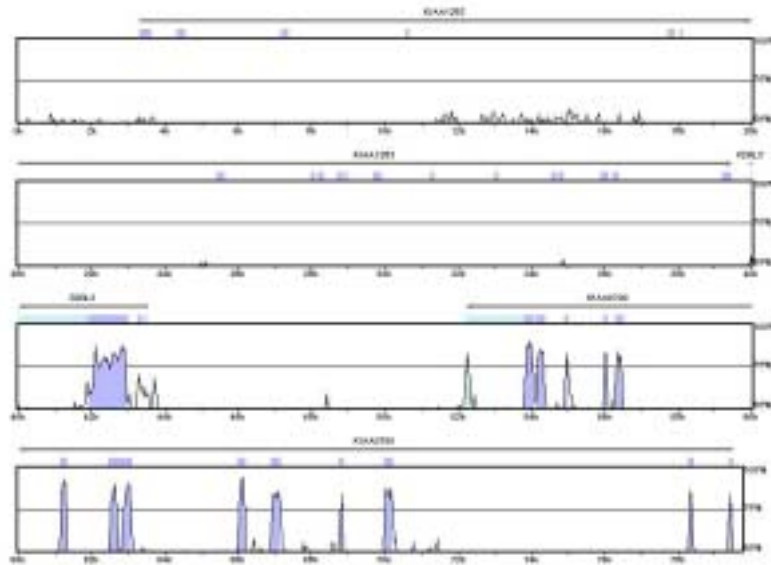
Preservation of
large (huge)
blocks of chromosome
over evolutionary time,
Preserving gene order

Last count:
Chimp 5 mln years: 1
Mouse 80 mln years: 183
Fugu fish 400-1000s ???



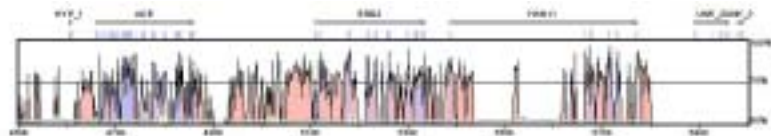
Known syntenic breakpoint F2RL3

Human
mouse

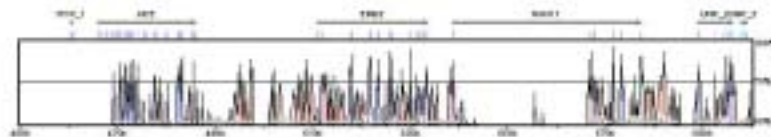


Second thing: proper distance from human

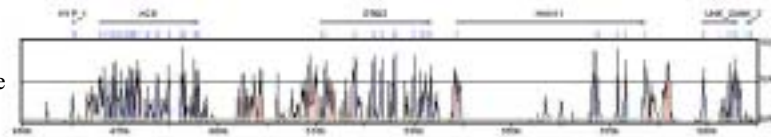
human-lemur



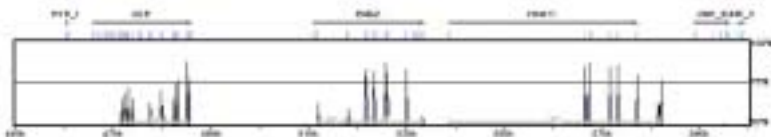
human-rabbit



human-mouse

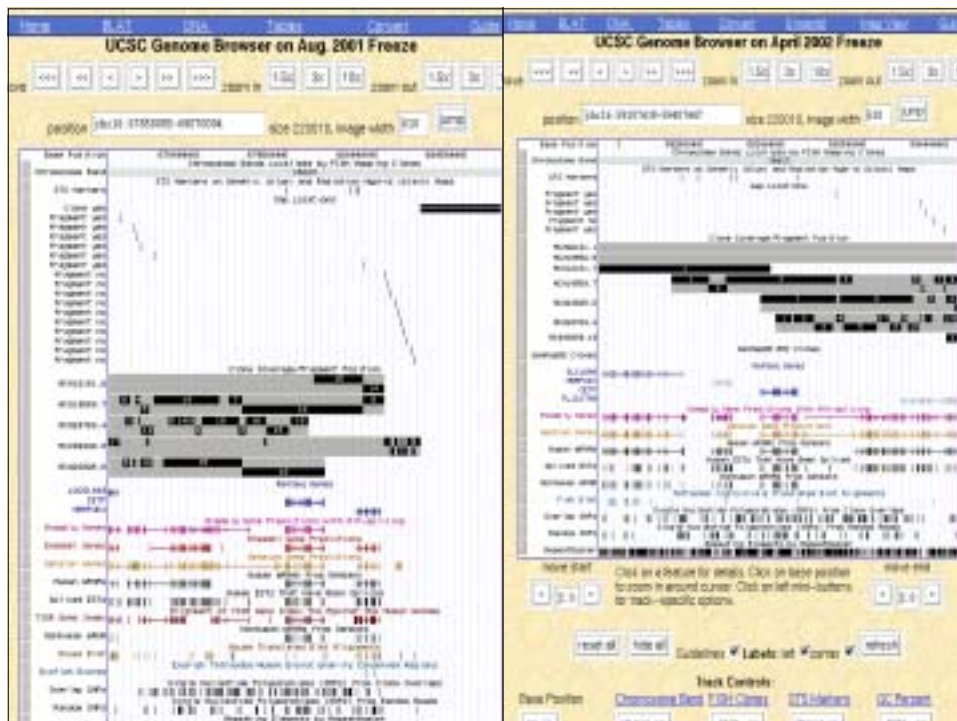
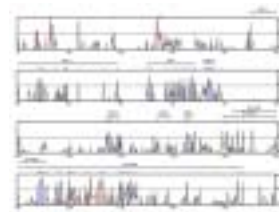


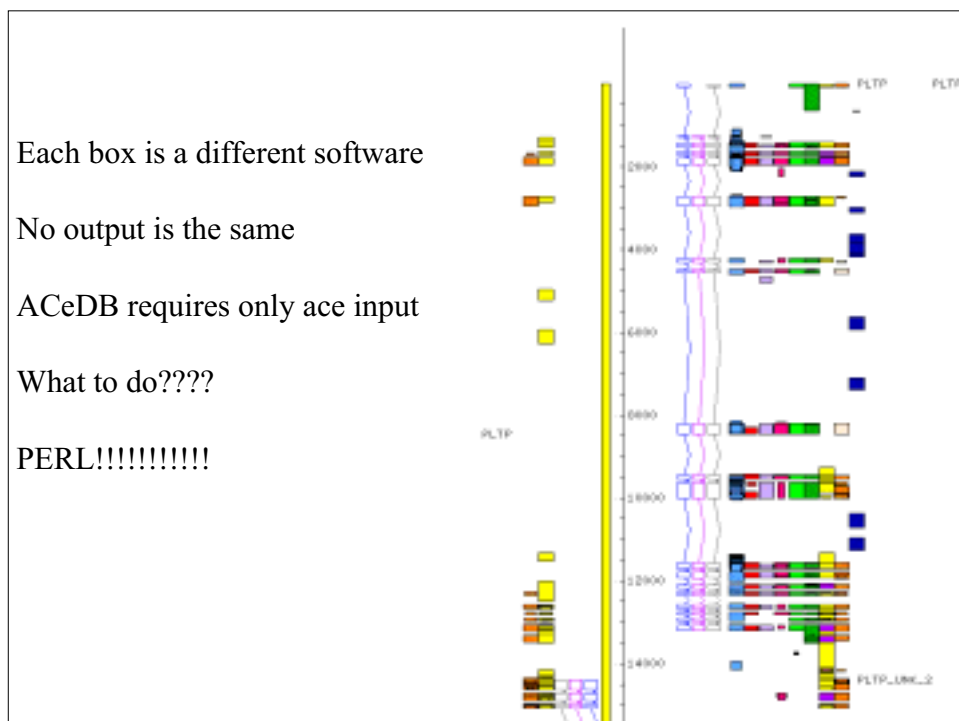
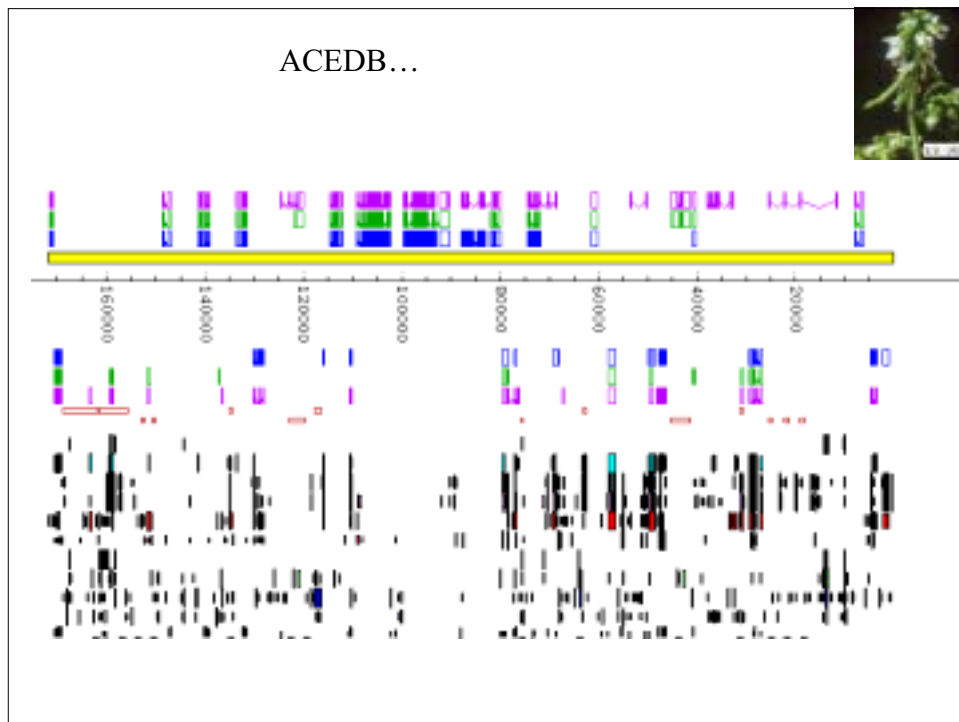
human-chick

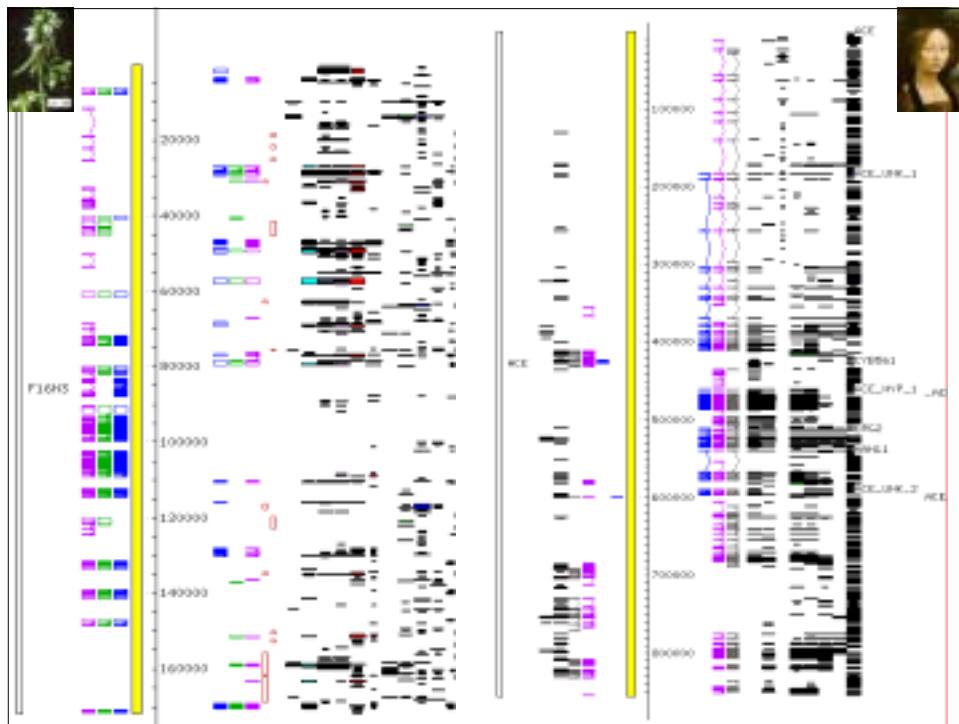


How do we make a vista?

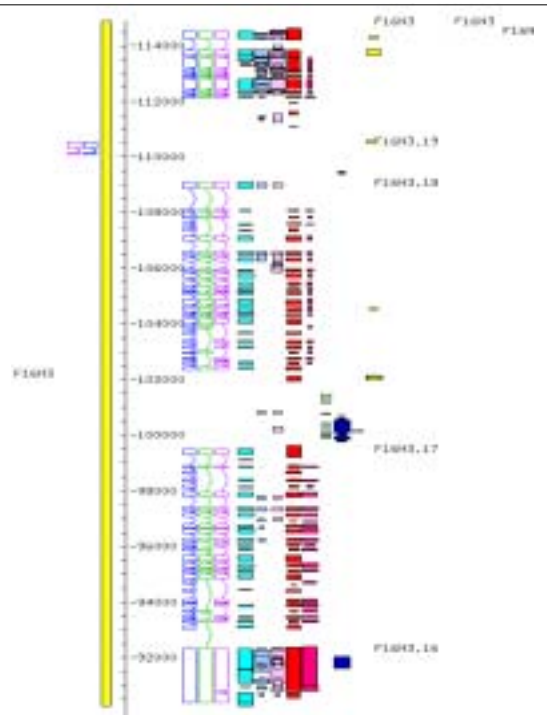
- Need *finished* human sequence
- Human sequence must be annotated
- Need to sequence and assemble second sequence
- Run alignment – get 1 vista
- Total time 1 to 2 weeks
- Godzilla browser- region in seconds
- Is what I do any better?



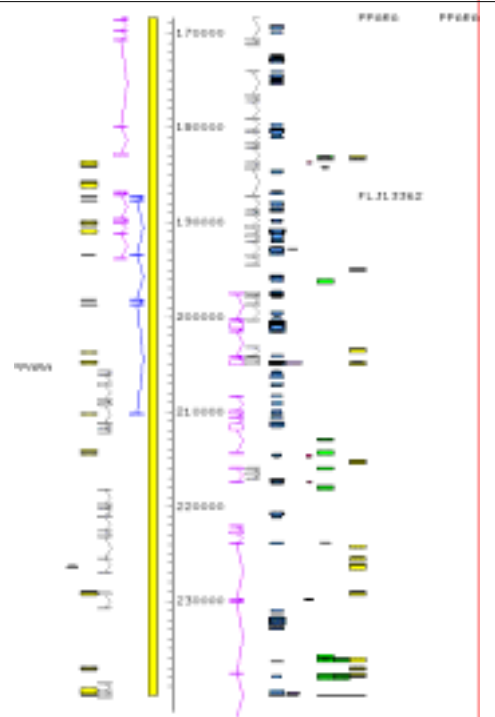




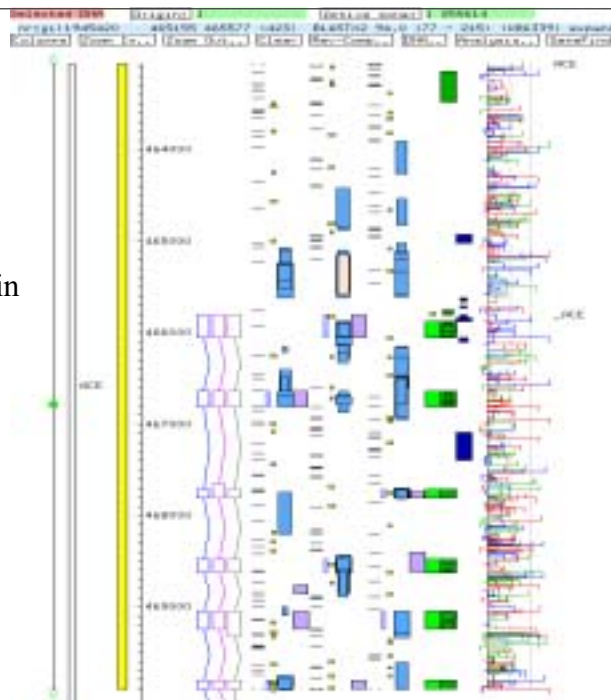
Arabidopsis
 No blast evidence
 But if it looks like a gene,
 It is a gene (Probably).



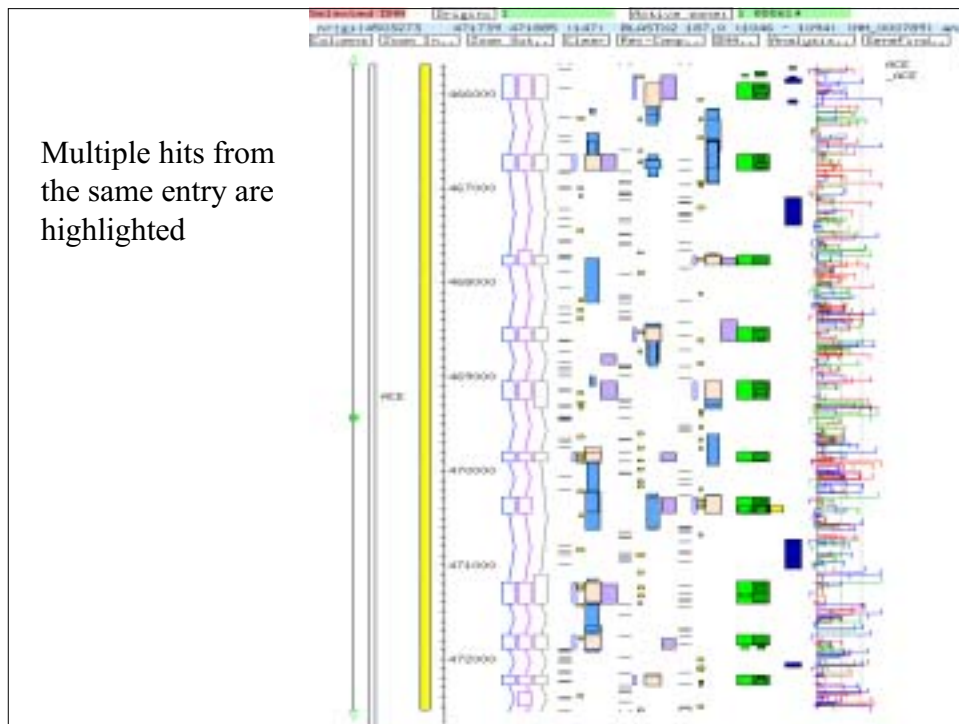
My First Human



Blastx
Sequence translated in
6 frames against the protein
database

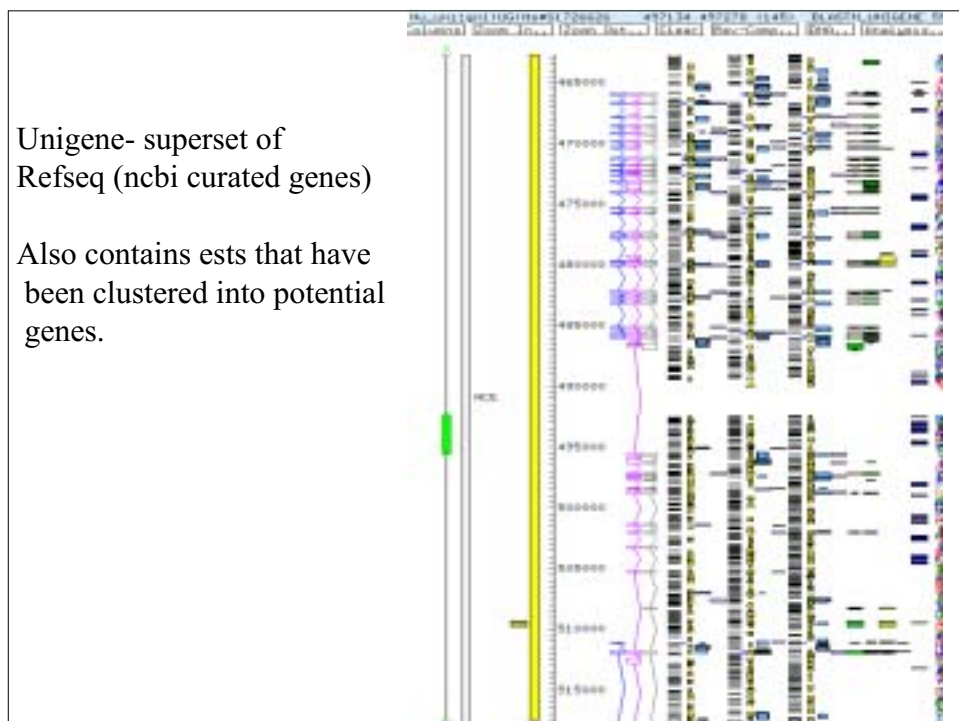


Multiple hits from the same entry are highlighted



Unigene- superset of Refseq (ncbi curated genes)

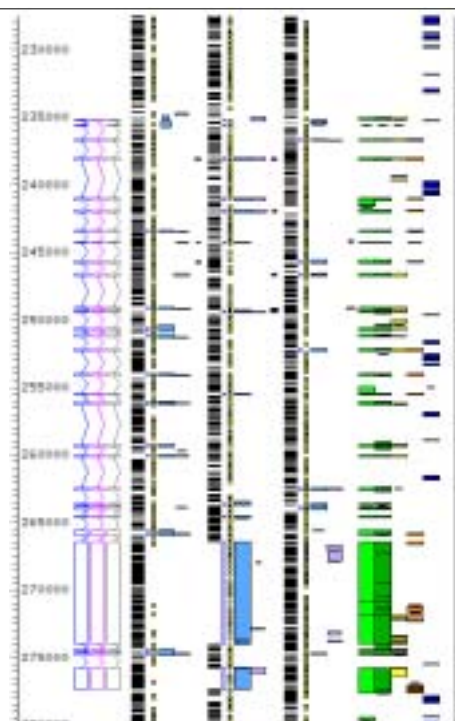
Also contains ests that have been clustered into potential genes.



This is not a gene



This is a gene



Start Here

NCBI LocusLink

Published Entrez BLAST OMIM Tipology Structure

Search LocusLink Display Ref Organism All

Query: APB [Go] [Clear]

Introduction

LocusLink provides a single query interface to curated sequence and descriptive information about genetic loci. It presents information on official nomenclature, aliases, sequence accessions, phenotypes, EC numbers, MIM numbers, UniGene clusters, homology, map locations, and related web sites.

Sequence accessions include a subset of GenBank accessions for a locus, as well as a new type, the NCBI Reference Sequence (RefSeq). RefSeq records are built according to the process detailed here. See the [Access](#) and [FAQ](#) pages for more information.

Data can be accessed by clicking one of the letters above to browse loci sorted by symbols, or by entering a query into the search form. Use of wild cards (*) is supported. Additional information and query tips are provided in the [help](#) documentation.

The current scope is fruit fly, human, human immunodeficiency virus type 1, mouse, rat, and zebrafish.

New Features

March, 2002: [MeSH](#) indexing staff are now providing [Gene22](#) (References into Function) data shown in the Function section of the LocusLink report. As part of the indexing process, new papers discussing the basic biology of a gene (function, structure, genetics) are being linked to the LocusLink report with a concise text summarizing the importance of the paper to the understanding of that gene and its products. Because of the [LinkOut](#) function, users retrieving articles in PubMed will also be able to navigate to LocusLink to learn more about the genes being referenced.

Feb 22, 2002: RefSeqs are now being generated for *Drosophila*, the zebrafish. NCBI RefSeqs for *Drosophila melanogaster* were first released in November, 2001. Current statistics for these genomes are available [here](#). We gratefully acknowledge the collaborations of [Ensembl](#) and [ZFIN](#).

Mv = map view

NCBI LocusLink

APB Index: Top of Page Nomenclature Overview Function Relationships Map RefSeq GenBank Links

LocusLink: Collaborators Download FAQ Help Statistics

RefSeq: About Download FAQ Statistics

Gene Ontology™:

Term	Evidence	Source	Pub
apb	NR	Proteome	
microsome	P	Proteome	pm
catalase	NR	Proteome	
apb-like protein	NR	Proteome	
soluble fraction	NR	Proteome	
signal transduction	NR	Proteome	
endoplasmic reticulum	NR	Proteome	
low-density lipoprotein	P	Proteome	pm

Other Ontologies:

Term	Evidence	Source	Pub
Soluble	NR	Proteome	pm
Intercellular transport	NR	Proteome	

Relationships

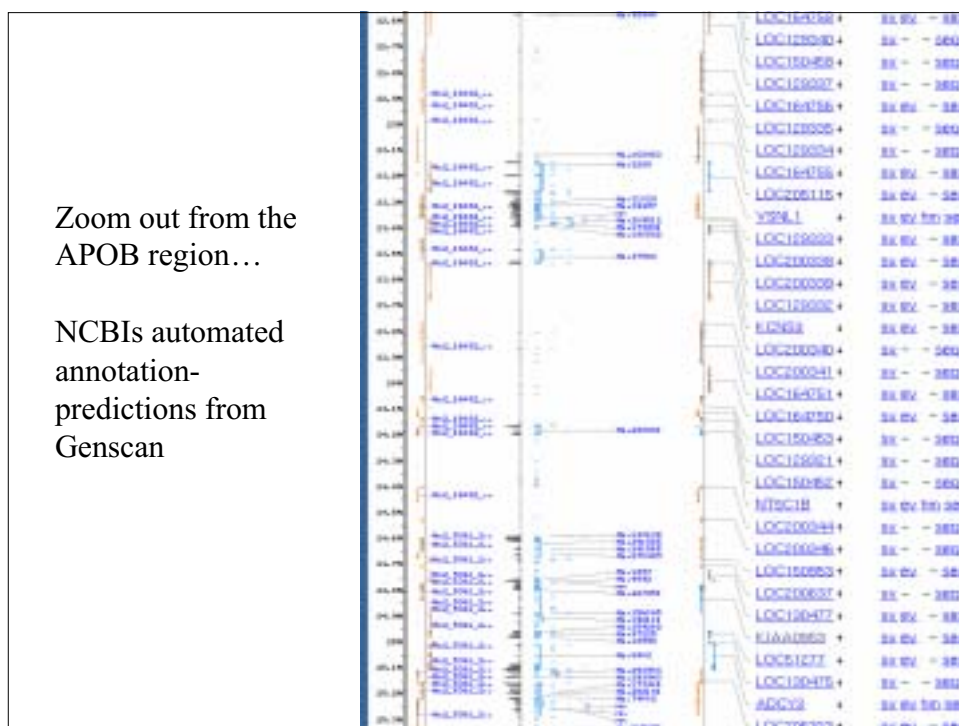
Mouse Homology Maps:

NCBI vs. MGD	UCSC vs. MGD	NCBI vs. EST-based RH Map	UCSC vs. EST-based RH Map	UCSC vs. Hudson et al.
12 2.00 cM	12 2.00 cM	12 112.01 cR	12 112.01 cR	12 18516.41 cR
A933	A933	A97111425	A97111425	A97111425
12 112.01 cR	12 112.01 cR	12 112.01 cR	12 112.01 cR	12 112.01 cR
A933	A933	A97111425	A97111425	A97111425
12 112.01 cR	12 112.01 cR	12 112.01 cR	12 112.01 cR	12 112.01 cR
A933	A933	A97111425	A97111425	A97111425

Map Information

Chromosome:	Cytogenetic:	Markers:
2	2p24-p29	
		HUGO
	Chr. 2	B93665
	Chr. 2	H50951
	Chr. 2	S0032670
	Chr. 2	SHGC-24681
	Chr. 2	O032588
	Chr. 2	O032588
	Chr. 2	O032588

NCBIs automated
annotation-
predictions from
Genscan



KIAA.../FLJ...
cdnas isolated
From Japan.
Protein product is
hypothetical

XM ...

Key for display of mRNAs aligning in this region:

[Map/View Evidence Viewer Help](#)

■ Genomic sequence (C)
■ model exons, single (M)
■ model exons, [overlapping](#) (M)
■ mRNA exons, single (G, R)
■ mRNA exons, [overlapping](#) (G, R)
 C = contig; M = model mRNA; R = RefSeq mRNA; G = GenBank; mRNA

EST density key (E):

■ 1 EST ■ 2-5 ESTs ■ 6-20 ESTs
■ 21-99 ESTs ■ >100 ESTs

29 exons and 1 gene found in this genomic region spanning 43445 bp.

[View graphic only.](#)

LocusLink:
<http://www.ncbi.nlm.nih.gov/LocusLink/>

LOC164914

Key for display of mRNAs aligning in this region:

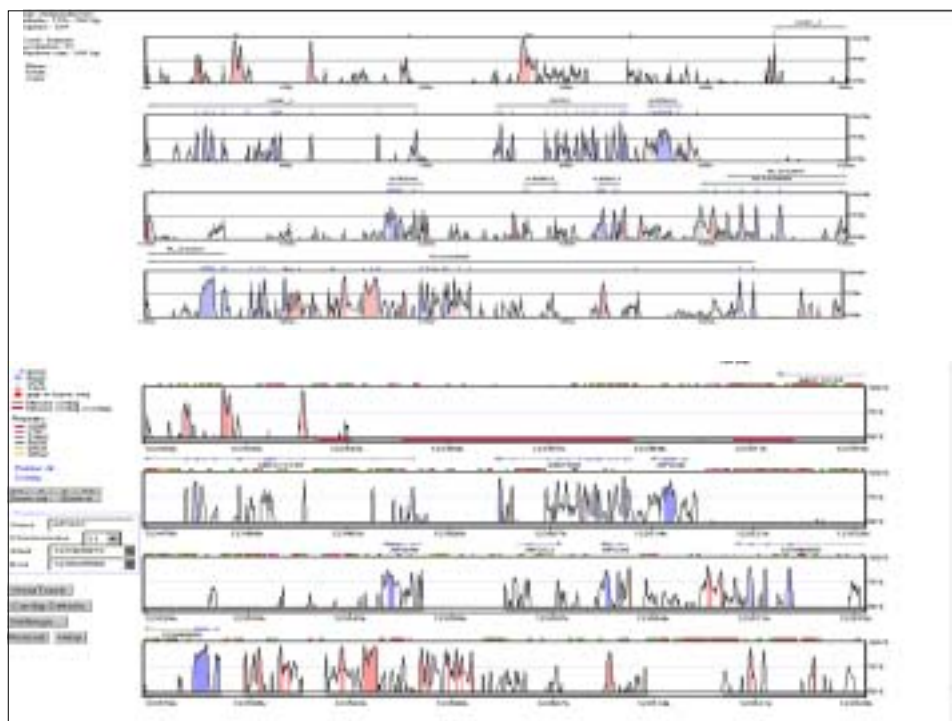
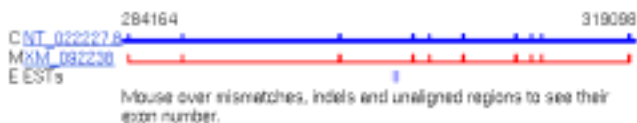
[Map/View](#)
[Evidence Viewer](#)
[Help](#)

■ Genomic sequence (C)
■ model exons, single (M) ■ mRNA exons, single (G, R)
■ model exons, overlapping (M) ■ mRNA exons, overlapping (G, R)
C = contig; M = model mRNA; R = RefSeq mRNA; G = GenBank mRNA

EST density key (E):

■ 1 EST ■ 2-5 ESTs ■ 6-20 ESTs
■ 21-99 ESTs ■ >100 ESTs

0 exon and 1 gene found in this genomic region spanning 34935 bp.
[View graphic only](#)



Assembly: for example Phrap

Overlapping reads



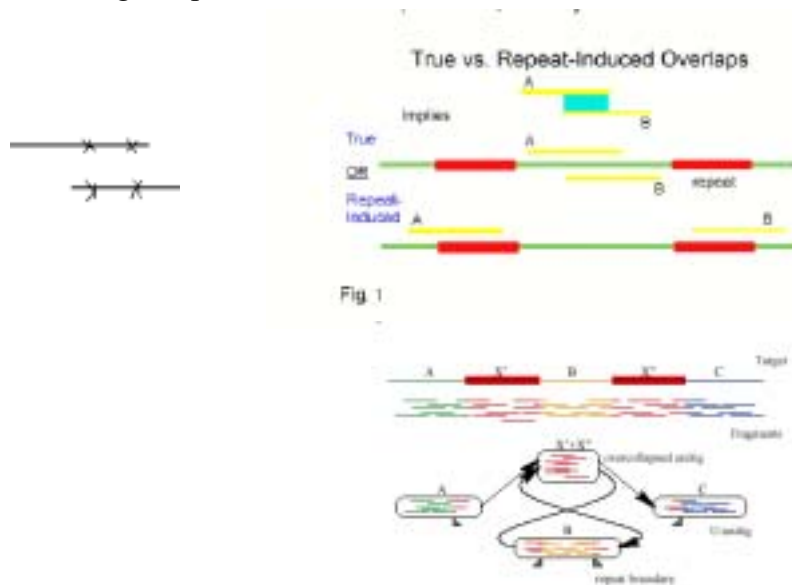
All groups of overlapping sequence
make up one contig

Larger structures called supercontigs
can be made by connecting the
paired-ends over the sequence gaps



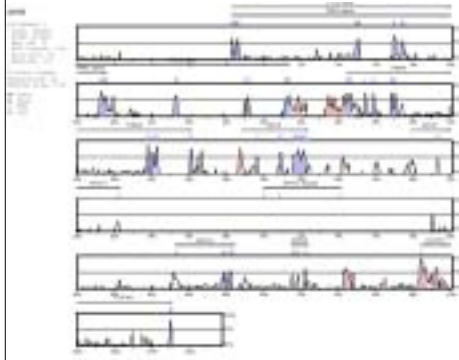
These are perfect overlaps.
The computer can figure these out easily.
Simple right?

Wrong. Repeats are confused with base call errors.

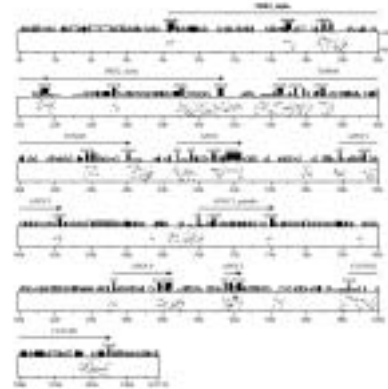


Alignment Method

- GLOBAL- AVID



- LOCAL- Pinmaker



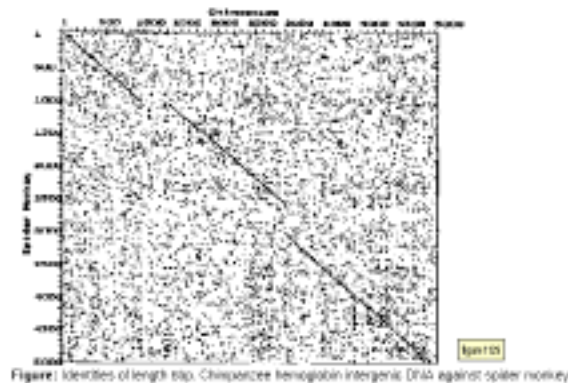
Both use “heuristics” to approximate either a “real” global or local alignment.

To compare 2 sequences, slide them past each other to see where they match maximally, allowing gaps

AGCCUCG AGCC UCG
CCGCCAUCG CCGCCAUCG = 5

Computers can represent this as a matrix:

	A	G	C	C	U	C	G
C		*	*		*		
C		*	*		*		
G	*						*
C		*	*		*		
C		*	*		*		
A	*						
U					*		
C		*	*		*		
G							*



Needleman-Wunsch in 5 minutes
The “real thing” global alignment

	M	P	R	C	L	C	Q	R	J	N	C	B	A
P			1										
B												1	
R				1				1					
C					1	1					1		
K													
C					1	1					1		
R								1					
N										1			
J									1				
C					1	1					1		
J									1				
A													1

The math says, take the max
of either:
the diagonal plus the match
the row one down plus the match
the column one down plus the match

Add the 4 plus one for the match = 5

	i												
	M	P	R	C	L	C	Q	R	J	N	C	B	A
P	0	1	0	0	0	0	0	0	0	0	0	0	0
B	0	0	1	1	1	1	1	1	1	1	1	2	1
R	0	0	2	1	1	1	1	2	1	1	1	1	2
C	0	0	1	3	2	3	2	2	2	2	3	2	2
K	0	0	1	2	3	3	3	3	3	3	3	3	3
C	0	0	1	3	3	4	3	3	3	3	4	3	3
R	0	0	2	2	3	3	4	5					
N													
J										1			
C				1	1						1		
J										1			
A													1

It looks like: $H_{ij} =$
 $\max(H_{i-1,j-1} + s(A_i, b_i)),$
 $\max(H_{i-k,j-1} + s(A_i, b_i)),$
 $\max(H_{i-1,j-k} + s(A_i, b_i)),$

You then look for the highest Score. It will always be in the Last row or column. Since you have kept track of where you came from, you can just “traceback” to the other end. This gives you a global alignment of two sequences, it is the maximal score for those two sequences, allowing all gaps.

	M	P	R	C	L	C	Q	R	J	N	C	B	A
P	0	1	1	0	0	0	0	0	0	0	0	0	0
B	0	0	1	1	1	1	1	1	1	1	1	2	1
R	0	0	2	1	1	1	1	2	1	1	1	1	2
C	0	0	1	3	2	3	2	2	2	2	3	2	2
K	0	0	1	2	3	3	3	3	3	3	3	3	3
C	0	0	1	3	3	4	3	3	3	3	4	3	3
R	0	0	2	2	3	3	4	5	4	4	4	4	4
M	0	0	1	2	3	3	4	4	5	5	5	5	5
J	0	0	1	2	3	3	4	4	6	5	6	6	6
C	0	0	1	3	3	4	4	4	5	5	7	6	6
J	0	0	1	2	3	3	4	4	6	6	6	7	7
A	0	0	1	2	3	3	4	4	5	5	6	7	8

MP-RCLQR-JNCBA
 | | | | | | |
 -PBCCKC-RNJ-CJA

Smith –Waterman: Local

Same algorithm, with one exception, you get a penalty when you move off the diagonal

This means that gaps are not free. For each square away, you are Penalized 1/3 in this case.

	C	A	G	C	C	B	C	G	C	U	U	A	G
A	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
A	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.1
U	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.1
G	0.0	0.0	1.0	0.5	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7	1.0
C	1.0	0.0	0.0	2.0	1.5	0.3	1.0	0.5	2.0	0.1	0.5	0.5	0.5
C	1.0	0.7	0.0	1.0	0.5	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7
A													
U													
U													
G													
A													
C													
G													

This means that you can get a negative score. If you get a negative score, it stays 0, and you stop.

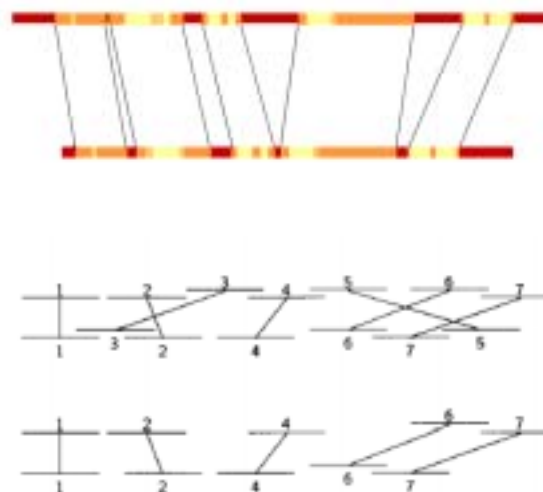
What happens is that you end up with a maximal score anywhere within the two sequences, you don't have to go to the ends.

	C	A	G	C	C	U	C	G	C	U	U	A	G
A	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
A	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
U	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7
G	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7	1.0
C	1.0	0.0	0.0	2.0	1.3	0.3	1.0	0.3	2.0	0.7	0.3	0.3	0.3
C	1.0	0.7	0.0	1.0	1.0	1.7	1.3	1.0	1.3	1.7	0.3	0.0	0.0
A	0.0	2.0	0.7	0.3	1.7	2.7	1.3	1.0	0.7	1.0	1.3	1.3	0.0
U	0.0	0.7	1.7	0.3	1.3	2.7	2.3	1.0	0.7	1.7	2.0	1.0	1.0
U	0.0	0.3	0.3	1.3	1.0	2.3	2.3	2.0	0.7	1.7	2.7	1.7	1.0
G	0.0	0.0	1.3	0.0	1.0	1.0	2.0	2.0	1.7	1.3	2.3	2.7	2.7
A	0.0	1.0	0.0	1.0	0.3	0.7	0.7	2.0	0.0	1.7	1.3	2.3	2.0
C	1.0	0.0	0.7	1.0	2.0	0.7	1.7	1.7	0.0	2.7	1.3	1.0	2.0
G	0.0	0.7	1.0	0.3	0.7	1.7	0.3	2.7	1.7	2.7	2.3	1.0	2.0
G	0.0	0.0	1.7	0.7	0.3	0.3	1.3	1.3	2.3	1.3	2.3	2.0	2.0

GCC-UCG

GCCAUUG

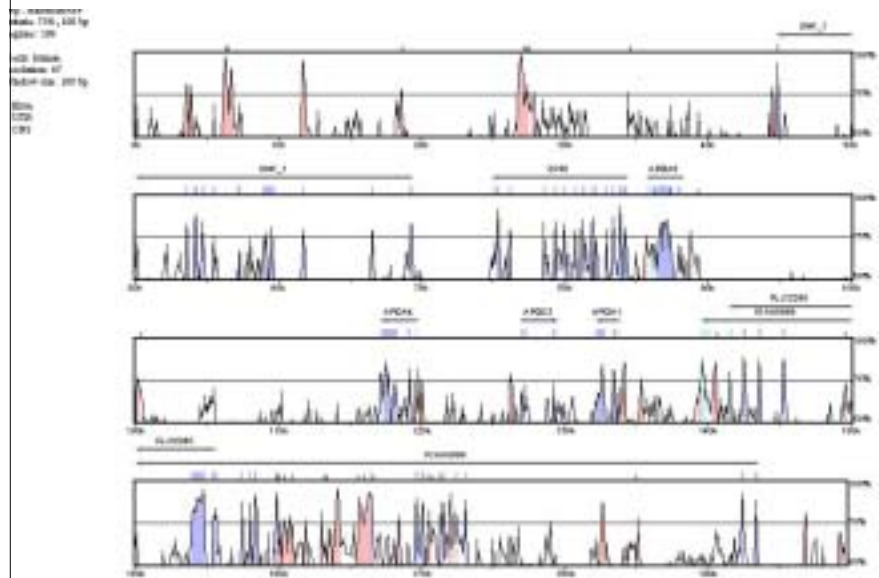
Global Heuristic- AVID



BLAT

[illegible]

Voila



Aknowledgements:

Eddy Rubin
Inna Dubchak
Jan-Fang Cheng
My friends in the Rubin lab